

THE SYNTHETIC-OVERSAMPLING METHOD: USING PHOTOMETRIC COLORS TO DISCOVER  
EXTREMELY METAL-POOR STARSA. A. MILLER<sup>1,2,3\*</sup>

DRAFT January 24, 2017

## ABSTRACT

Extremely metal-poor (EMP) stars ( $[\text{Fe}/\text{H}] \leq -3.0$  dex) provide a unique window into understanding the first generation of stars and early chemical enrichment of the Universe. EMP stars are exceptionally rare, however, and the relatively small number of confirmed discoveries limits our ability to exploit these near-field probes of the first  $\sim 500$  Myr after the Big Bang. Here, a new method to photometrically estimate  $[\text{Fe}/\text{H}]$  from only broadband photometric colors is presented. I show that the method, which utilizes machine-learning algorithms and a training set of  $\sim 170,000$  stars with spectroscopically measured  $[\text{Fe}/\text{H}]$ , produces a typical scatter of  $\sim 0.29$  dex. This performance is similar to what is achievable via low-resolution spectroscopy, and outperforms other photometric techniques, while also being more general. I further show that a slight alteration to the model, wherein synthetic EMP stars are added to the training set, yields the robust identification of EMP candidates. In particular, this synthetic-oversampling method recovers  $\sim 20\%$  of the EMP stars in the training set, at a precision of  $\sim 0.05$ . Furthermore,  $\sim 65\%$  of the false positives from the model are very metal-poor stars ( $[\text{Fe}/\text{H}] \leq -2.0$  dex). The synthetic-oversampling method is biased towards the discovery of warm ( $\sim \text{F-type}$ ) stars, a consequence of the targeting bias from the SDSS/SEGUE survey. This EMP selection method represents a significant improvement over alternative broadband optical selection techniques. The models are applied to  $>12$  million stars, with an expected yield of  $\sim 600$  new EMP stars, which promises to open new avenues for exploring the early universe.

*Subject headings:* methods: data analysis – methods: statistical – stars: general – stars: statistics – stars: fundamental parameters – surveys

## 1. INTRODUCTION

Understanding the origins of structure on all scales, from the largest filaments containing galaxy clusters, to the smallest biological lifeforms that inhabit planets orbiting stars within the galaxies in those clusters, is arguably the main tenet of astronomy. The recent proliferation of wide-field surveys aims to study these problems, and a vast array of related questions, by generating large statistical samples that capture the diversity of different objects throughout the Universe. A challenge for these surveys, however, is that more data is not equivalent to better data. While the Large Synoptic Survey Telescope (LSST; Ivezić et al. 2008a) will eventually dwarf all other ground-based, wide-field optical surveys, the data deluge from LSST demands the development of superior algorithmic techniques. These methods must be capable of capturing and exploiting complex information from current and future data streams.

Data-driven methods, such as machine-learning algorithms, provide an intriguing solution to these challenges. These models are extremely flexible and have the ability to ascertain complex, non-linear interactions within the data. In brief, machine-learning models use a training set, a collection of sources with known *labels*, such as a classification or physical property, to derive a mapping between those labels and *features*, measured properties of the sources in the training set. Once the mapping

is learned, this knowledge can be applied to new, unlabeled data. With spectroscopic resources already in short supply, a major challenge is deriving labels that are traditionally determined from spectroscopic measurements, e.g., redshift or metallicity, from photometric observations alone. The importance of solutions to this problem will be amplified during the LSST-era, when more than 20 billion sources will be photometrically detected (Ivezić et al. 2008a). The majority of these sources will not be amenable to spectroscopic observations, even with thirty-meter class telescopes.

Almost from the time it was realized that metal-rich stars produce less light in the blue optical than their metal-poor counterparts (Schwarzschild et al. 1955), efforts have been made to photometrically estimate stellar metallicities (e.g., ultraviolet-excess technique; Wallerstein 1962). The most successful efforts to date use narrowband and mediumband filters, designed to be sensitive to metallicity dependent absorption lines in the stellar spectrum. The most prominent technique uses the *uvby $\beta$*  Strömgren filters (see Strömgren 1966 for a review), which have been demonstrated to produce  $[\text{Fe}/\text{H}]$ <sup>5</sup> measurements with a scatter of  $\sim 0.1$  dex relative to spectroscopic observations for FG stars (Nordström et al. 2004). For isolated groups of stars (clusters, galaxies), if there is a single stellar population (i.e., the were born during a single episode of star formation) and the distance is well known, then  $[\text{Fe}/\text{H}]$  estimates can be made

<sup>1</sup> Jet Propulsion Laboratory, 4800 Oak Grove Drive, MS 169-506, Pasadena, CA 91109, USA

<sup>2</sup> California Institute of Technology, Pasadena, CA 91125, USA

<sup>3</sup> Hubble Fellow

\* E-mail: amiller@astro.caltech.edu

<sup>5</sup> Throughout this paper  $[\text{Fe}/\text{H}]$  is used as a proxy for metallicity, where  $[\text{Fe}/\text{H}]$  is defined as  $\log(N_{\text{Fe}}/N_{\text{H}})_* - \log(N_{\text{Fe}}/N_{\text{H}})_\odot$ , where  $N_{\text{Fe}}$  and  $N_{\text{H}}$  are the total number of iron and hydrogen atoms, respectively.

using a single photometric color and isochrone fitting. Interestingly, [Lianou et al. \(2011\)](#) find that isochrone fitting performs poorly relative to spectroscopic methods when the single stellar population assumption is violated. Modern wide-field surveys, such as the Sloan Digital Sky Survey (SDSS; [York et al. 2000](#)) or LSST, primarily observe field stars with broadband filters. The age and distance to any field star is highly uncertain, meaning methods that use the SDSS filters are at a significant disadvantage relative to the Strömgren filters or isochrone fitting. Nevertheless, when careful selections are made to limit samples to FG stars, broadband photometric estimates of  $[\text{Fe}/\text{H}]$  can be made with a typical scatter of  $\sim 0.2\text{--}0.3$  dex (e.g., [Ivezić et al. 2008b](#); [Bond et al. 2010](#)). Achieving this precision requires the use of the  $u$ -band (see [An et al. 2009](#)).

Stellar atmospheres retain the composition of the gas from which the star forms: as the universe becomes enriched with metals over time, so do newly formed stars. Thus, stellar metallicity measurements can serve as a proxy for stellar age (though the scatter in these relations is large, see [Soderblom 2010](#) for a review). Stars with very small metal abundances, known as extremely metal-poor (EMP) stars ( $[\text{Fe}/\text{H}] \leq -3.0$  dex), are relics from the early universe that provide unique insight to the nature of the first generation of stars. In particular, stars with  $M_*/M_\odot \lesssim 0.8$  have not had sufficient time, within the age of the universe, to undergo significant post-main-sequence evolutionary changes and remain on, or close to, the main sequence. Therefore, the atmospheres of EMP stars retain information on the initial mass function of Population III stars, the diversity and nucleosynthetic yield of the first supernovae, the early chemical enrichment of the universe, and the formation of the first galaxies (for recent reviews on EMP stars see [Beers & Christlieb 2005](#); [Frebel & Norris 2015](#)). As a result, considerable efforts have been made to identify EMP stars in the Milky Way halo in order to understand the nature of the Galaxy in the first  $\sim 500$  Myr after the Big Bang.

Traditionally, candidate EMP stars are identified via objective-prism or low-resolution-spectroscopic surveys, and later confirmed via high-resolution spectroscopy. The HK Survey of [Beers et al. \(1985, 1992\)](#) identified EMP candidates from stars with weak Ca II  $K$  absorption. Several groups have utilized objective-prism observations from the Hamburg/ESO survey to identify EMP candidates and confirm bonafide EMP stars with high-resolution spectroscopy (e.g., [Cohen et al. 2004](#); [Frebel et al. 2006](#); [Christlieb et al. 2008](#)). Recently, SDSS, and in particular the SDSS-II sub-survey known as the Sloan Extension for Galactic Understanding (SEGUE; [Yanny et al. 2009](#)), have identified hundreds of EMP candidates from low-resolution spectra. Many of these candidates have been confirmed with high-resolution observations (e.g., [Aoki et al. 2013](#)). Additional follow-up is on-going for all of these surveys, and more EMP discoveries can be expected.

Early evidence of the utility of the ultraviolet-excess technique suggested that the relation saturated for very metal-poor (VMP) stars ( $[\text{Fe}/\text{H}] \leq -2.0$  dex), and this result has been seemingly confirmed with modern survey data (e.g., [Bond et al. 2010](#)). As a result, there have been virtually no studies on the utility of identifying EMP stars from broadband photometric colors alone.

Recently, [Schlaufman & Casey \(2014\)](#) developed a technique that exploits the significant near-infrared molecular absorption of metal-rich stars to identify candidate EMP stars. Using data from the Two Micron All Sky Survey (2MASS; [Skrutskie et al. 2006](#)) and the *Wide-field Infrared Survey Explorer* (WISE; [Wright et al. 2010](#)), [Schlaufman & Casey](#) identify bright ( $V < 14$  mag) EMP candidates, of which a small handful, corresponding to an efficiency of a few percent, have been confirmed via their initial follow-up spectroscopy. Additionally, the SkyMapper Telescope is poised to discover a large bounty of EMP stars by combining observations from the broadband *ugriz* filters with a custom narrow filter centered on the Ca II  $K$  line ([Keller et al. 2007](#)). The use of this narrow filter is extremely efficient for the discovery of EMP stars, and the early returns from SkyMapper include the confirmation of 41 EMP stars via high-resolution spectroscopy ([Jacobson et al. 2015](#)). The unique filter combination has also led to the discovery of the most iron-poor star known ([Keller et al. 2014](#)). SkyMapper follow-up is still ongoing, and estimates of the discovery efficiency using their narrow band filter are currently not available (though 41 of the 122 EMP candidates studied in [Jacobson et al. 2015](#) were confirmed, suggesting an efficiency of  $\sim 1/3$ ). Nevertheless, this survey likely represents the premier method for uncovering southern sky EMP stars in the near future.

Here, a new technique to estimate  $[\text{Fe}/\text{H}]$  from only broadband *ugriz* filters is presented. The method utilizes machine-learning algorithms and is trained using a sample of  $\sim 170,000$  stars with precise photometric observations and spectroscopic determinations of  $[\text{Fe}/\text{H}]$  from SDSS. It is demonstrated that the method is superior to other photometric  $[\text{Fe}/\text{H}]$  techniques. Furthermore, the method can be slightly altered, via the inclusion of synthetic EMP stars in the training set, to be suitable for the discovery of EMP stars. This final model enables the first-ever identification of EMP stars from broadband-optical filters alone.

## 2. THE SPECTROSCOPIC SAMPLE

Machine-learning models require a training set: a collection of sources with known labels. Once the mapping between features and labels is learned, the model can be applied to newly-observed, unlabeled sources for which only features are known. The construction of the training set and choice of machine-learning algorithm are essential steps for constructing a model that produces accurate predictions. Furthermore, as is the case for all data-driven approaches, the training set must be representative of the population of unlabeled sources or the model predictions will be unreliable. This is a major challenge for many astronomical surveys: typically, new surveys probe fainter populations than those present in previously studied well-labeled samples (see e.g., [Richards et al. 2012](#)). As detailed in §6, significant care is taken to ensure that the models developed here are only applied to the subset of field stars that is extremely similar to training set stars.

### 2.1. SDSS Spectroscopic Measurements of $[\text{Fe}/\text{H}]$

SDSS is an optical, wide-field survey that has produced *ugriz* imaging of  $> 14,500$  deg<sup>2</sup> and collected spectra of  $> 850,000$  stars (several million spectra of extragalactic

targets have also been obtained; Alam et al. 2015). With  $> 250,000,000$  stars without spectroscopic observations, the SDSS dataset is ideal for the construction of the model: the large reservoir of spectroscopically observed stars will ensure a robust training set, yet there remains a significant pool of sources to search for candidate EMP stars.

All SDSS optical stellar spectra are analyzed via the automated Segue Stellar Parameters Pipeline (SSPP; for full details on the SSPP see Lee et al. 2008a,b; Allende Prieto et al. 2008). Briefly, the SSPP determines  $T_{\text{eff}}$ ,  $\log g$ , and  $[\text{Fe}/\text{H}]$  for stellar sources using multiple parameter estimation methods (e.g., neural networks, synthetic spectral matching, Ca II K line index technique, etc.). The individual measurements of  $T_{\text{eff}}$ ,  $\log g$ , and  $[\text{Fe}/\text{H}]$  are then robustly combined to provide final adopted values, and their corresponding uncertainties. For high signal-to-noise ratio (SNR) spectra with  $4500 \text{ K} \leq T_{\text{eff}} \leq 7000 \text{ K}$ , the SSPP determines  $T_{\text{eff}}$ ,  $\log g$ , and  $[\text{Fe}/\text{H}]$  with typical uncertainties of 157 K, 0.29 dex, and 0.24 dex, respectively. In addition to estimates of these parameters, the SSPP provides processing flags for sources where the parameter estimates are no good, such as white dwarfs or M stars.

## 2.2. Training Set Selection Criteria

Photometric colors and spectroscopic  $[\text{Fe}/\text{H}]$  measurements for the training set sources are selected from SDSS data release 10 (DR10; Ahn et al. 2014), which includes the most recent version of the SSPP. In total there are 427,225 sources with  $[\text{Fe}/\text{H}]$  measurements in DR10, however, this set is further pruned to avoid systematic biases and ensure a high-quality training set. The selection criteria are designed to select sources with the most reliable photometric and spectroscopic measurements. It is important to note, each of these criteria can be applied to the  $\sim 2.6 \times 10^8$  SDSS stars with no spectroscopic observations, ensuring that these choices do not introduce a significant bias in the final model predictions.

Poor, or missing, photometric measurements will corrupt the fidelity of the machine-learning models, thus, the first restrictions placed on the training set are photometric. The following photometric properties can all be retrieved from the `PhotoObjAll` table in the SDSS DR10 database. The first requirement for inclusion in the training set is a detection in each of the *ugriz* bands, equivalent to `psfMag_f`  $> 0$ , where **f** is the SDSS filter [427,177; in this and the next paragraph the number of sources remaining in the training set following each constraint will be given in brackets]. Good calibration in each filter, `calibstatus_f`  $= 1$ , where, again, **f** is the filter [420,575], and a non-flagged photometric measurement, i.e. `clean`  $= 1$  [399,646], are also required. Sources fainter than 19.5 mag in the *g*-band are excluded, `psfMag_g`  $\leq 19.5$  [390,741]. Finally, sources with large photometric uncertainties are excluded, as these will result in a noisy mapping between colors and  $[\text{Fe}/\text{H}]$ . Sources with `psfMagErr_u`  $\geq 0.04$  mag, or `psfMagErr_h`  $\geq 0.03$  mag where **h** is any of the *griz* filters are excluded [240,614].

Spectroscopic properties are retrieved from the DR10 `sppParams` table. The SSPP is most reliable for stars over a restricted range in  $T_{\text{eff}}$ ,  $4500 \text{ K} \leq \text{TEFFADOP} \leq 7000 \text{ K}$  (Lee et al. 2008a), thus, stars outside this range

are excluded [216,593].<sup>6</sup> Furthermore, only stars with at least two individual measurements of  $[\text{Fe}/\text{H}]$  are included, `FEHADOPN`  $\geq 2$  [209,163], as some of the individual SSPP methods for  $[\text{Fe}/\text{H}]$  measurements do not perform well over the full range of observed metallicities (see Schlesinger et al. 2012). Requiring two  $[\text{Fe}/\text{H}]$  measurements significantly reduces the likelihood of a pathologically incorrect  $[\text{Fe}/\text{H}]$  measurement. Finally, only sources with the following SSPP flags are included: `nnnnn`, `nnngn`, or `nnnGn`, which correspond to normal stars, stars with a slight G-band feature, and stars with a potentially strong G-band feature, respectively [197,059]. Sources with any other combination of flags likely have unreliable  $[\text{Fe}/\text{H}]$  measurements and are unsuitable for this study (Y. S. Lee, private communication). Finally, for stars with multiple spectra only the highest SNR spectrum is retained in the training set [170,610].

These 170,610 stars form the training set, and the SSPP measured  $[\text{Fe}/\text{H}]$  values form the labels for the model. Prior to computing stellar colors, the observed brightness in each filter is de-reddened using the Schlafly & Finkbeiner (2011) recalibration of the Schlegel et al. (1998), hereafter SFD98, dust maps. The reddening corrected photometric colors,  $(u-g)_0$ ,  $(g-r)_0$ ,  $(g-i)_0$ , and  $(g-z)_0$ , constitute the full feature set for the model.

This reddening correction introduces some uncertainty into the model, however, the majority of SDSS observations are at high galactic latitudes, where extinction, and its corresponding correction, is small. The SFD98 dust maps measure the total Galactic reddening along a given sightline, meaning these corrections are equivalent to assuming the stars in this study reside outside the Milky Way. This assumption is clearly false, and maximally correcting for reddening in this way may result in some stars with colors that are too blue. Nevertheless, it is assumed that the bias from this overcorrection is small, especially because SDSS observations focused on low-extinction sight lines [ $> 85\%$  (92%) of the training set has  $A_r \leq 0.2$  (0.3) mag]. Furthermore, with a bright limit of  $g \approx 14$  mag and a sample composed primarily of FG stars, the majority of stars in this study are  $\gtrsim 1$  kpc away, meaning the adopted reddening correction is reasonable. This assumption is further corroborated by the generally good performance of the model (see §4).

Reddening corrections become extremely problematic near the Galactic plane ( $|b| \lesssim 10^\circ$ ), where the SFD98 maps are unreliable and extinction is very patchy. As a result, the methods presented here will provide unreliable  $[\text{Fe}/\text{H}]$  estimates near the plane, unless superior extinction estimates to individual stars are developed. Further discussion of potential biases introduced by the reddening correction is provided in the conclusions (§7).

The scope of the training set is shown in Figure 1, which displays several summary statistics on a  $u-g$ ,  $g-r$  color-color (CC) diagram. The training set covers the full extent of the stellar locus, while spanning metallicities from EMP stars to metal-rich stars ( $[\text{Fe}/\text{H}] > 0.0$  dex). Figure 1(a) demonstrates that  $[\text{Fe}/\text{H}]$  can be readily determined from photometric colors.

<sup>6</sup> Strictly speaking,  $T_{\text{eff}}$  cannot be determined for stars with only photometric measurements. However, the photometric relation for  $T_{\text{eff}}$  provided in Pinsonneault et al. (2012) will enable the removal of stars that are too hot or too cool for the machine-learning model.



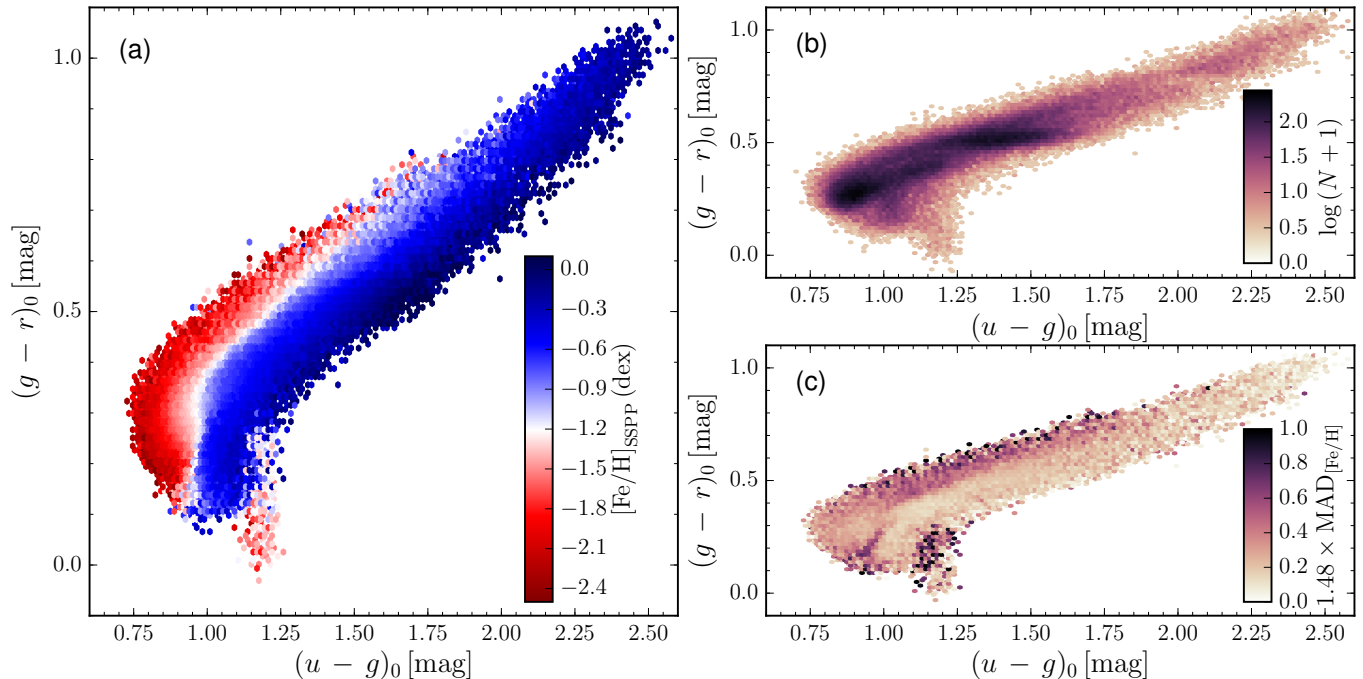


FIG. 1.— Summary of the training set shown in the  $(u - g)_0$ ,  $(g - r)_0$  CC diagram. Each plot shows summary statistics for the stars located within individual pixels, which are  $\approx 0.01$  mag on a side. Only pixels with  $\geq 2$  stars are shown. (a): the median  $[\text{Fe}/\text{H}]$  per pixel. Note that at constant  $(g - r)_0$  color, roughly corresponding to constant  $T_{\text{eff}}$ , the  $(u - g)_0$  color provides an excellent diagnostic of  $[\text{Fe}/\text{H}]$ . (b): the total number of stars per pixel. Machine-learning methods typically perform best in regions where there is ample training data. The strong over-density with  $0.48 < (g - r)_0 < 0.55$  is due to the SDSS emphasis on targeting G stars, while the over-density at  $\approx (0.9, 0.3)$  is due to the targeting emphasis on F-turnoff-like stars (see e.g., Yanny et al. 2009). (c): the scatter in  $[\text{Fe}/\text{H}]$ , as measured by  $1.48$  multiplied by the median absolute deviation (MAD), per pixel. Pixels with small scatter represent locations where the machine-learning model will be most accurate.

### 3. MACHINE LEARNING MODELS

In addition to building a robust and representative training set, the choice of machine-learning algorithm is essential for the construction of a useful model.<sup>7</sup> Below a brief overview of the three different algorithms utilized in this study is provided.

#### 3.1. *K*-nearest neighbors

*K*-nearest neighbors (KNN) regression identifies the *K* training-set sources that are closest to the newly observed source in feature space. For the new source, the predicted value of the target variable is simply the mean from the *K* neighbors. An advantage of KNN regression is that it is simple, and the model results are easy to interpret. In this study, KNN regression is performed using the Python `scikit-learn` implementation of the algorithm (Pedregosa et al. 2011). Scaling factors are applied to each individual color so that the re-scaled features have a sample mean of zero and sample variance unity prior to performing KNN regression.

#### 3.2. Random forest

Random forest (RF) methods utilize the aggregation of multiple decision trees to assign a final classification or regression value to newly observed sources (Breiman 2001). RF makes use of bagging (see Breiman 1996), wherein bootstrap samples of the training set are used to construct each of the  $N_{\text{tree}}$  total trees in the forest. As each tree in the forest is constructed, only a random

subset of  $m_{\text{try}}$  features is selected from the full feature set as a potential splitting criterion at each node of the tree. The use of bagging and  $m_{\text{try}}$  random features reduces the variance of the final model predictions relative to single decision-tree models, providing low-bias, low-variance results. The final RF predictions are determined by averaging the predictions for a new source from each of the  $N_{\text{tree}}$  individual trees. Furthermore, the RF algorithm is fast, each of the trees can be constructed independently and thus in parallel, and relatively easy to interpret. RF models have recently become highly popular as an application for astronomical data sets due to their relative insensitivity to noisy or meaningless features (e.g., Brink et al. 2013; Miller et al. 2015), and their invariant response to even highly non-gaussian feature distributions (e.g., Dubath et al. 2011; Richards et al. 2011). This study utilizes the Python `scikit-learn` implementation of the RF algorithm (Pedregosa et al. 2011).

#### 3.3. Support vector machines

Support vector machines (SVMs; Boser et al. 1992; Cortes & Vapnik 1995) are learning models that project the features from the training set into a high- or infinite-dimension space. SVMs then find a linear hyperplane with the maximal margin separating the two groups of sources, in the case of classification. These methods can be generalized to regression problems (Drucker et al. 1997), where the hyperplane must produce predictions on the training set that are within a given threshold of their true values. For this study a non-linear radial basis function is used to perform SVM regression, which is

<sup>7</sup> For a general overview of machine learning, we refer the interested reader to Hastie et al. (2009).

TABLE 1  
TEST SET PREDICTIONS

Model	RMSE (dex)	CER	Train Time <sup>a</sup> (s)
KNN	0.297	0.028	0.1
RF	0.297	0.027	72.4
SVM	0.294	0.027	728.4

NOTE. — Models: KNN – k-nearest neighbors, RF – random forest, SVM – support vector machines. All models have been optimized using 10-fold cross validation and a grid search of their respective tuning parameters. The results shown here reflect the average of 5 separate models, and the model-to-model scatter is small.

<sup>a</sup> Average time required to train the model on 110,000 sources using dual-socket, 8-core, 2.66 GHz Intel Sandy Bridge CPUs with 64 GB of memory.

implemented using the LIBSVM software package (Chang & Lin 2011). For the SVM model, the individual colors are re-scaled so that the minimum and maximum values of the features are 0 and 1, respectively.

#### 4. REGRESSION MODEL RESULTS

##### 4.1. Comparison of the Three Regression Models

To determine which of the three models from §3 best generalizes to new data, the 170,610 spectroscopic sources were separated into a training set containing 110,000 sources and a validation set with 60,610 sources. The models are optimized via a grid search over the relevant tuning parameters using 10-fold cross validation (CV) performed on the 110,000 source training set. The parameters that minimize the root-mean-square error (RMSE):

$$\text{RMSE} = \left[ \frac{1}{n} \sum_i^n (y_i - x_i)^2 \right]^{1/2},$$

where  $n$  is the total number of sources in the training set,  $y_i$  is the model prediction of  $[\text{Fe}/\text{H}]$  for the  $i^{\text{th}}$  source, and  $x_i$  is the  $[\text{Fe}/\text{H}]$  spectroscopic value for the  $i^{\text{th}}$  source. Small changes in the optimal tuning parameters do not significantly alter the CV RMSE. The optimal models were applied to the 60,610 source validation set, with the results summarized in Table 1 and the final predictions shown in Figure 2.

The panels in Figure 2 show that the KNN, RF, and SVM models all produce similar predictions for  $[\text{Fe}/\text{H}]$  based on photometric colors. Formally, the SVM model produces the best predictions with RMSE = 0.2943 dex, which is  $\sim 1\%$  better than the KNN and RF models. The SVM model also has the lowest catastrophic error rate, CER, defined as the fraction of sources where the predicted and spectroscopic values of  $[\text{Fe}/\text{H}]$  differ by  $\geq 0.75$  dex. Again, while the SVM model has the best performance the difference between the three is small  $\sim 1\text{--}3\%$ . The residuals, shown in the bottom panel of Figure 2, are also similar for the three models. For stars with  $[\text{Fe}/\text{H}]_{\text{SSPP}}$  between  $\sim 0$  and  $-2$ , corresponding to the vast majority of stars in the Galaxy (Schlesinger et al. 2012), the models exhibit virtually unbiased predictions with small scatter. There is a systematic bias for stars with  $[\text{Fe}/\text{H}]_{\text{SSPP}} \lesssim -2$  or  $[\text{Fe}/\text{H}]_{\text{SSPP}} \gtrsim 0$ , which have over- and under-predicted values of  $[\text{Fe}/\text{H}]$ , respectively.

##### 4.2. Understanding the Regression-Model Bias

As a measure of the overall bias of each model, the Pearson  $r$  correlation coefficient is measured for the residuals as a function of spectroscopic  $[\text{Fe}/\text{H}]$  values. An unbiased model would show little to no correlation,  $|r| \approx 0$ . Models with  $|r| \rightarrow 1$  show a strong correlation between the residuals and  $[\text{Fe}/\text{H}]$ , indicating significant bias in the final model predictions. The SVM model has the smallest  $|r|$ , meaning it has the smallest bias of the three machine-learning models.

The correlation between the residuals and either  $T_{\text{eff}}$ , the individual photometric colors, or  $\log g$ , is significantly weaker than the correlation between the residuals and  $[\text{Fe}/\text{H}]_{\text{SSPP}}$ . Using the Fisher transformation of the Pearson  $r$  coefficient, the correlation of the residuals with each of the photometric colors,  $T_{\text{eff}}$ , and  $\log g$  is significantly smaller, probability  $P \ll 0.0001$ , than the correlation with  $[\text{Fe}/\text{H}]$ .

Thus, the systematic biases seen in Figure 2 are most likely the result of alternative effects. There are two systematic effects that play a role in this bias: (i) regression to the (sample) mean, and (ii) regression dilution bias. Non-parametric, data-driven regression models often produce predictions biased towards the sample mean. This effect can most easily be illustrated for KNN models. Consider the most metal-poor star in the validation set, which has  $[\text{Fe}/\text{H}] = -3.68$  dex, at best, the KNN prediction for this source would be the mean  $[\text{Fe}/\text{H}]$  of the 60 most metal-poor stars<sup>8</sup> from the training set, which is equal to  $-3.38$  dex. This represents the best case scenario, if the nearest neighbors for this EMP star include any that are not the least metal-poor in the training set the model-predicted  $[\text{Fe}/\text{H}]$  will be biased even further from the true value. The models are also susceptible to bias due to the uncertainties associated with the photometric colors and spectroscopic  $[\text{Fe}/\text{H}]$  measurements. Noisy features and target variables lead to a flattening of the regression slope, an effect known as regression dilution bias (Frost & Thompson 2000). This bias could be improved in the future with more precise color measurements and superior spectroscopic determinations of  $[\text{Fe}/\text{H}]$ , though it may be prohibitively expensive to obtain these observations. Further discussion of these two types of bias can be found in Miller et al. (2015).

Physical effects may also be responsible for the systematic over-prediction of  $[\text{Fe}/\text{H}]$  for VMP stars. As metals are removed from a stellar atmosphere, the absorption lines present become weaker and weaker. Eventually, at some critical metallicity,  $Z_{\text{crit}}$ , the lines will become so weak that they can no longer be detected via broadband-photometric colors. This means photometric-metallicity techniques eventually saturate, and assign all stars with  $Z < Z_{\text{crit}}$  the same  $[\text{Fe}/\text{H}]$ . If  $Z_{\text{crit}}$  occurs at  $[\text{Fe}/\text{H}] \approx -2.0$  dex, then this would naturally explain some of the bias seen in Figure 2. The photometric technique presented in Bond et al. (2010) shows a similar saturation for stars with  $[\text{Fe}/\text{H}] \lesssim -2.0$  dex. Nevertheless, in §5 it is shown that EMP stars can be recovered using broadband optical colors, meaning the saturation of photometric metallicity is not solely responsible for the biased VMP star predictions.

<sup>8</sup>  $K = 60$  for the optimized KNN model.

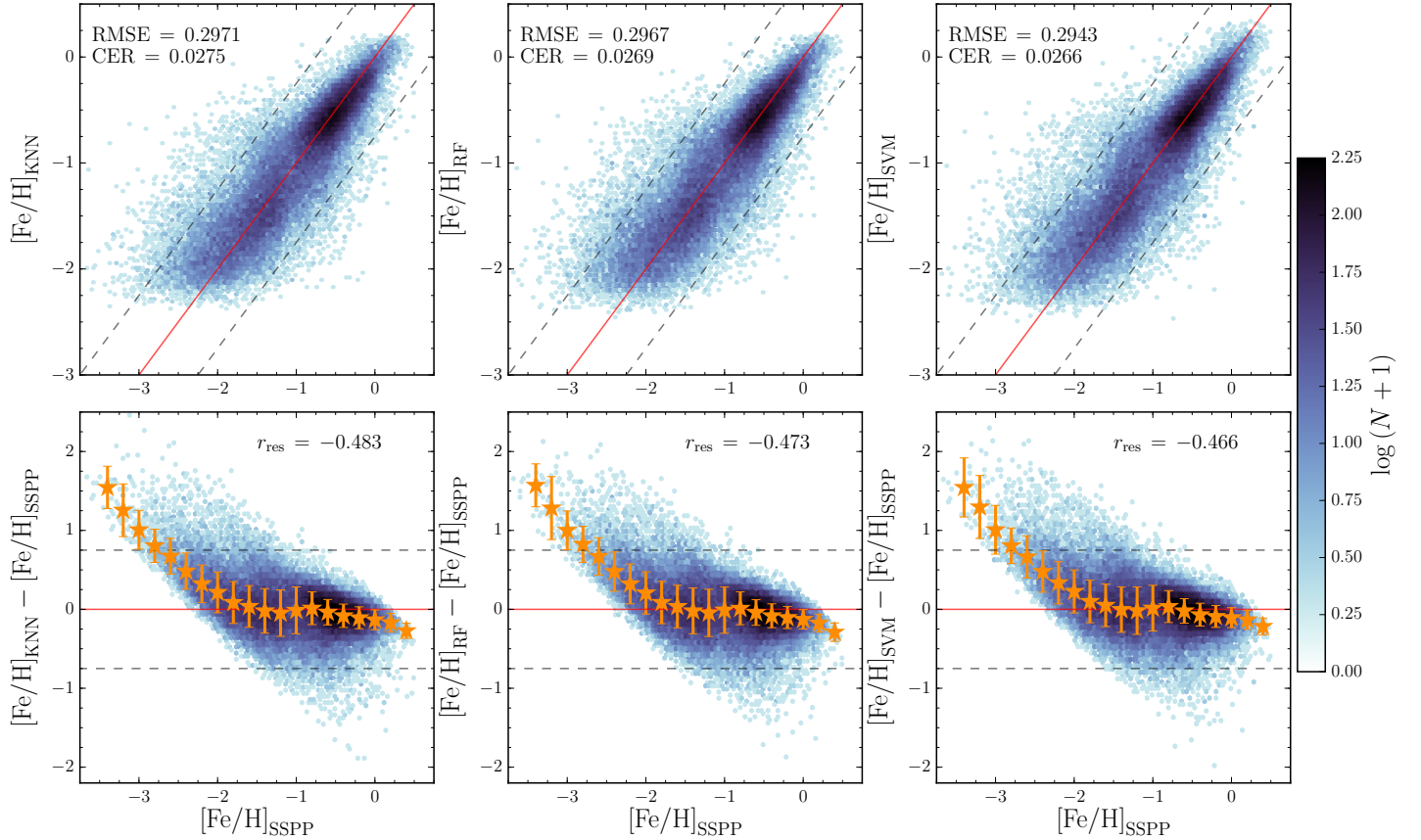


FIG. 2.— Regression results for the optimized KNN, RF, and SVM models shown, respectively, in the three columns from left to right. *Top*: Density plot showing the number of sources in each pixel on the predicted  $[\text{Fe}/\text{H}]$  vs. SSPP  $[\text{Fe}/\text{H}]$  plane. Pixels are  $\sim 0.05$  dex on a side. The solid, red line shows the relation for perfect 1:1 regression, while the dashed, grey lines show the boundaries for catastrophic errors,  $\pm 0.75$  dex. Sources located outside the grey lines are considered catastrophic outliers. The SVM model has the smallest RMSE and CER. *Bottom*: Residuals from the models (shown above), with the density of sources shown in each pixel. The orange stars show the median residual value in bins of width 0.2 dex. The associated errorbars show the scatter in each bin:  $1.48 \times \text{MAD}$ , an outlier-resistant and robust measure of the scatter.  $r_{\text{res}}$  is the Pearson  $r$  correlation coefficient for the residuals as a function of  $[\text{Fe}/\text{H}]_{\text{SSPP}}$ .  $r_{\text{res}}$  values close to zero indicate little bias in the model predictions. The SVM model produces the least biased estimates of  $[\text{Fe}/\text{H}]$ .

#### 4.3. Comparison to spectra

With an RMSE scatter of  $\sim 0.29$  dex, the SVM model produces predictions of  $[\text{Fe}/\text{H}]$  that are similar to those from low-resolution spectra. The SSPP provides  $[\text{Fe}/\text{H}]$  measurements with a typical uncertainty of  $\sim 0.24$  dex (Lee et al. 2008a), though this precision is limited to stars with high SNR ( $\gtrsim 50$ ). To better facilitate the comparison between this study and the SSPP results, we compare the scatter between the 77 stars in this study that are also part of the SSPP high-resolution validation set (see Tables 3 and 4 of Allende Prieto et al. 2008).<sup>9</sup> Adopting the  $[\text{Fe}/\text{H}]$  values measured from the high-resolution spectra as ground-truth, then the SSPP has an RMSE = 0.37 dex, while the SVM model has an RMSE = 0.44. With the caveat that this comparison is based on a small number of stars, this suggests that the SSPP performs  $\sim 17\%$  better than the SVM model. These RMSE values are significantly larger than those reported in Lee et al. (2008a), because the 77 stars in common between this study and the sample in Allende Prieto et al. (2008) primarily excludes relatively metal-rich stars. The analysis presented

in Allende Prieto et al. (2008) consists of two samples: a relatively metal-rich sample (median  $[\text{Fe}/\text{H}] \approx -0.5$  dex) observed with the Hobby Eberly Telescope (HET), and a relatively metal-poor sample (median  $[\text{Fe}/\text{H}] \approx -2.0$  dex) observed with the Keck and Subaru telescopes. The initial study compares 81 stars observed with HET and 44 stars observed by Keck and Subaru, while the sample in common with this study retains only 42 HET stars and 35 from Keck and Subaru. The scatter between the SSPP and the high-resolution measurements from Keck and Subaru (0.41 dex; see Table 6 of Allende Prieto et al. 2008) is significantly worse than the scatter for stars observed with HET (0.12 dex). Both the SSPP and the SVM model perform better on relatively metal-rich stars, thus, the preferential exclusion of these stars in the HET sample would lead to a corresponding increase in the RMSE.

#### 4.4. Comparison to other photometric methods

With photometrically observed stars in SDSS outnumbering spectroscopically observed stars by nearly a factor of  $\sim 10^3$ , there have been many efforts focused on determining photometric metallicity estimates from broadband SDSS colors. In Kerekes et al. (2013), a KNN method is used to predict  $[\text{Fe}/\text{H}]$  with an RMSE  $\approx 0.32$

<sup>9</sup> Most of these 77 stars are in the 110,000 star training set. Thus, the SVM predictions here are from 10-fold CV to avoid an overlap between the training and test sets.



dex for stars with  $15 \text{ mag} < g < 17 \text{ mag}$ , and 0.41 dex for stars with  $18 \text{ mag} < g < 19 \text{ mag}$ . The sample in the Kerekes et al. study places no restrictions on the quality of the photometric or spectroscopic observations. Thus, stars that raised SSPP flags or have large photometric uncertainties are likely driving the significantly larger RMSE from that model.

Multi-dimensional polynomial fits to the median  $[\text{Fe}/\text{H}]$  in  $0.02 \text{ mag}^2$  bins in the  $(u - g)_0$ ,  $(g - r)_0$  plane are used to determine photometric metallicities in Ivezić et al. (2008b). This method is later updated in Bond et al. (2010), where SSPP values from DR7 replace the less accurate values from DR6, which were used in the Ivezić et al. study. The fit presented in Bond et al. (2010) produces a typical RMSE  $\sim 0.2$  dex for metal-rich stars and  $\sim 0.3$  dex for metal-poor stars. These values cannot be directly compared to those presented in §4.1, however, as the samples used in both the Ivezić et al. and Bond et al. studies placed more stringent cuts on the training set than those employed here. In particular, those studies included only sources with  $0.2 < (g - r)_0 < 0.6$ , so as to focus on F/G stars. If the same selection criteria from Ivezić et al. (2008b) are applied to the validation set from this study, 35,377 of the 60,610 stars remain. The RMSE for those stars is  $\sim 0.26$  dex for the SVM model and  $\sim 0.32$  dex for the photometric model presented in Bond et al. (2010).<sup>10</sup> Thus, the SVM model presented in this study represents an  $\sim 18\%$  improvement in the scatter relative to the polynomial-fitting method presented in Ivezić et al. (2008b) and Bond et al. (2010).

#### 5. MODEL ALTERATIONS TO EMPHASIZE THE SELECTION OF EMP STARS

While the regression models presented in §4 perform well for the vast majority of field stars, the strong biases for VMP stars make it difficult to identify EMP stars. The discovery of EMP stars can be cast as a classification problem where all EMP stars belong to one class and all other stars, with  $[\text{Fe}/\text{H}] > -3.0$ , form the other class. For the 170,610 stars in this study, 256 are EMP stars. Thus, there is a significant class imbalance between the EMP and non-EMP stars. Typically, machine-learning classification algorithms are built to maximize the overall accuracy of predictions. A classifier that predicts all stars belong to the majority non-EMP class would have an accuracy of 99.8%. For most machine-learning models this accuracy would be stunning. This masks the failure of the model for its most interesting task: identifying new EMP stars. Following some adjustments to the training set, however, EMP stars can be reliably recovered.

##### 5.1. Dealing with Class Imbalance: Upsampling and Downsampling

Many classification problems deal with imbalance, wherein at least one class represents a very small minority of sources. It is often the case, however, that the minority class represents the target of interest: identifying additional instances of these rare events is the motivation for model construction. Minimizing the overall classification error rate means special attention is not paid to the minority class and these sources are disproportionately misclassified. The consequences range from mildly

annoying, e.g., spam email bypassing filters to reach an inbox, to extremely serious, e.g., in the medical profession.

There are two general approaches for dealing with class imbalance. One approach is to manually adjust the imbalance in the training set by randomly downsampling the majority class or oversampling the minority class, or using a combination of the two. The other is to use cost-sensitive learning, where the cost for misclassification of the minority class is higher than the cost for misclassifying members of the majority. Most efforts focus on some form of the sampling technique, with downsampling approaches typically outperforming oversampling (see e.g., Chen et al. 2004). A downside to downsampling is that information is being removed from the classifier, while strict oversampling will always be fundamentally limited by the fact that no truly new instances of the minority class have been added to the classifier.

Many researchers have found that over-sampling the minority class with replacement does not significantly improve minority-class recognition (e.g., Ling & Li 1998). As a result, Chawla et al. (2002) developed the synthetic minority over-sampling technique (SMOTE), wherein synthetic members of the minority class are generated to reduce the class imbalance. In short, synthetic members are generated by fitting a KNN model to the minority class. For each minority-class source in the training set, one of the  $k$ -nearest neighbors is selected at random, and a synthetic member of the minority class is generated by selecting a random point along the feature vector connecting the source and its neighbor. This process is then repeated to achieve the desired amount of oversampling. While examining a variety of classic class-imbalance problems, SMOTE outperforms over-sampling, while the combination of SMOTE and downsampling performs better than both downsampling and cost-sensitive learning methods (Chawla et al. 2002; Chawla et al. 2003).

In Chen et al. (2004), two methods, which leverage both the sampling and cost-based approaches, are explored to improve the performance of RF on imbalanced data. The first approach, which they refer to as balanced random forest, uses a bootstrap sample of the minority class as well as an equal number of majority class members selected randomly with replacement to initiate each tree in the forest. Thus, the minority and majority classes are equally balanced over the classifier. The other method, which they refer to as weighted random forest, places a stronger penalty on misclassifying the minority class by weighting the samples when selecting splitting criteria at each node within individual trees and also weighting the final vote in the terminal nodes of each tree. Using multiple different performance measures, both methods show improvements relative to other techniques, including SMOTE plus down-sampling, over a variety of different problems (Chen et al. 2004).

##### 5.2. Improving Minority Class Recognition with Synthetic EMP Stars

Initial tests of the three methods described above, SMOTE, balanced RF, and weighted RF, showed no significant improvement in the recovery of EMP stars. This is likely the case due to the extreme imbalance for the problem at hand: the minority class constitutes less

<sup>10</sup> See their Equation A1, which is an update to Equation 4 presented in Ivezić et al. 2008b.

than 0.2% of the training set, which is significantly less than the datasets tested in Chawla et al. (2002); Chawla et al. (2003) and Chen et al. (2004). Instead, a SMOTE-inspired approach, which generates synthetic EMP stars in a different manner, is developed.

A zoom-in on the  $(u - g)_0$ ,  $(g - r)_0$  CC diagram is shown in Figure 3 with the location of the EMP stars in the training set highlighted. While there is a relatively tight cluster of EMP stars on the blue edge of the stellar locus, approximately centered at (0.8, 0.2), this roughly coincides with the highest density location of non-EMP stars. Furthermore, over half the EMP stars form a loose sequence along the upper portion of the stellar locus. Thus, SMOTE, which generates synthetic samples between nearest neighbors while ignoring any underlying structure, is liable to create synthetic EMP stars that lie off the relatively well defined sequence. Instead, a different approach, which I refer to as the synthetic-oversampling method, is adopted: new EMP stars are generated by resampling the photometric colors within the reported uncertainties from SDSS. In practice, the procedure is straightforward: EMP stars are selected randomly with replacement from the training set. The photometric measurements for each of the *ugriz* filters are then adjusted via a random number selected from a normal distribution with mean zero and standard deviation equal to the SDSS-measured photometric uncertainty in the respective filter. Colors for the synthetic stars are computed, and the SSPP [Fe/H] measurement for the original star is assigned to the synthetic star. Finally, the user may specify how many synthetic EMP stars are generated and included in the training set.

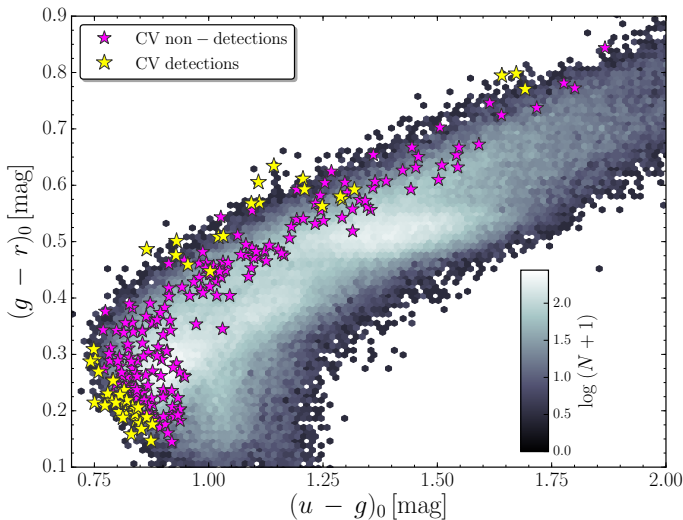


FIG. 3.—  $(u - g)_0$ ,  $(g - r)_0$  CC diagram showing the density of sources in the training set. The location of EMP stars in the training set is highlighted. Notice that the EMP stars form a relatively tight cluster that is parallel to the main stellar locus. EMP stars that are detected in CV via the synthetic-oversampling method are shown in yellow, while non-detections are shown in magenta. The synthetic-oversampling method is biased towards recovering those stars on the extremes of the sample distribution (see also §5.3).

As previously noted, the overall accuracy of a classifier is a poor measure of performance when trying to identify minority-class members in extremely imbalanced problems. Instead of focusing on the true positive rate and

false positive rate, respectively, I aim to simultaneously maximize the *precision* and *recall* of the model, which are defined as:

$$\begin{aligned} \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \end{aligned} \quad (1)$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.<sup>11</sup> The *precision* is a measure of how many erroneous measurements are required to recover a new member of the minority class, while the *recall* is a measure of the fraction of the minority class that is actually recovered. Ideally, a model would produce a *precision* and *recall*  $\approx 1$ , however, in practice, one must adopt a candidate decision threshold that offers a trade off between these two desirable features.

The SVM model is adopted to search for EMP stars due to its superior performance in the [Fe/H] regression problem discussed in §4. An SVM regression model is used, instead of an SVM classification model, so that candidates may be ranked by their likelihood of belonging to the EMP class. Thus, unlike a classification model where a single hard boundary between classes is determined, the class boundary from the regression model can be varied across different values of [Fe/H] to determine the optimal trade off between *precision* and *recall*. Given the rarity of EMP stars, the model is optimized via cross validation over the entire 170,610 source training set, rather than splitting the data into a training and validation set as was done in §4. The SVM tuning parameters are optimized via three different instances of 10-fold cross validation to maximize the *recall* at a *precision* of 0.05, which is adopted as the figure of merit (FoM). This FoM corresponds to only one in every 20 EMP candidates identified by the model being a genuine EMP star. While this performance seems relatively poor, it represents a dramatic improvement over previous broadband photometric techniques to identify EMP stars.

The results from optimized models with differing amounts of downsampling and synthetic oversampling are summarized in Table 2. As a baseline for the increase in performance downsampling and synthetic oversampling provide, the results for the full training set, with no synthetic over-sampling or downsampling, are also included. In addition to the FoM, Table 2 also includes other measures of model performance for imbalanced problems, including: the receiver operating characteristic (ROC) area under the curve (AUC), the *precision-recall* AUC, and the *F*-measure, defined as:

$$F\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

For Table 2 the *F*-measure is determined using a classification boundary of  $[\text{Fe}/\text{H}]_{\text{SVM}} = -3.0$  dex, where sources

<sup>11</sup> Note that there are many different nomenclatures throughout the machine-learning literature for the terms defined in Eqn. 1. *recall* is most commonly referred to as the true positive rate (TPR), though it can also be referred to as the sensitivity, hit rate, or completeness depending on the context. I adopt the convention of referring to this as the *recall* as this is only discussed relative to the *precision*. The *precision* of a model is sometimes referred to as the positive predictive value or purity.



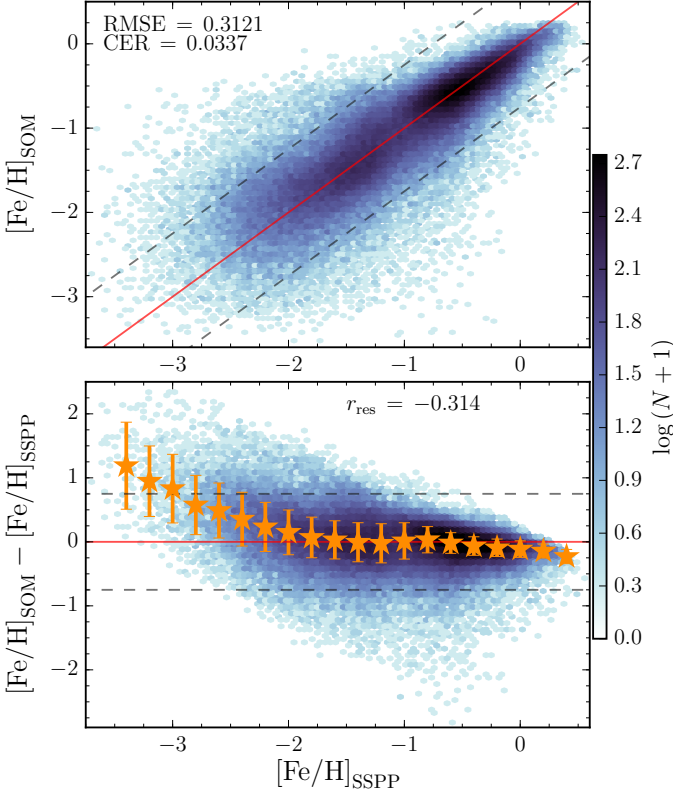


FIG. 4.— 10-fold cross-validated results for the full 170,610 star training set using the synthetic-oversampling model with 4000% oversampling and no downsampling. *Top*: Density plot showing the number of sources in each pixel on the predicted  $[\text{Fe}/\text{H}]$  vs. SSPP  $[\text{Fe}/\text{H}]$  plane. *Bottom*: Residuals from the model, with the density of sources shown in each pixel. The pixel scale, solid and dashed lines, and orange stars in the top and bottom plots are the same as in Figure 2. The synthetic-oversampling model produces less-biased predictions, with worse RMSE and CER, than the regression models in §4.

with a predicted  $[\text{Fe}/\text{H}]$  below this value are considered EMP candidates. The ROC curve traces the trade off between the true positive rate and the true negative rate as a function of classification decision boundaries. The ROC AUC measures the area beneath the ROC curve and is used to evaluate the overall performance of a classification model. The closer the ROC AUC is to 1, the better the model, though note that this metric does not consider *precision*. The *precision-recall* AUC measures the overall performance of a classifier on imbalanced data. Again, the closer this value is to 1, the better the model.

The FoM for the baseline model, which uses the entire training set with no synthetic oversampling, is 0.109. From Table 2, it is clear that the synthetic oversampling technique provides a significant improvement over the baseline model, with virtually all oversampled models showing a  $\sim 100\%$  increase in the FoM. Furthermore, it is clear that the precise choice of the degree of over- and downsampling does not have a strong effect on the final model predictions. With the exception of the model featuring 4000% oversampling and 25% downsampling, the largest difference in the FoM for oversampled models is  $\sim 10\%$ . Thus, I conclude that the use of a synthetically-oversampled minority class improves the efficiency of EMP star discovery.

Predictions of  $[\text{Fe}/\text{H}]$  from the synthetic-oversampling

model that features 4000% oversampling and no downsampling are shown in Figure 4. Relative to the SVM-regression model, there are far more stars with predicted  $[\text{Fe}/\text{H}] \leq -2.5$  dex. The synthetic-oversampling model is also less biased, as measured by the Pearson  $r$  coefficient. The overall performance of the model, as measured by the RMSE, is  $\sim 6\%$  worse than the SVM-regression model and the CER is  $\sim 27\%$  higher than the SVM-regression model. Each model has its relative strengths and weaknesses, and the ultimate choice of model should be driven by a user’s science goals. In particular, the synthetic-oversampling model is designed to identify EMP stars, while the SVM-regression model is designed to provide the most accurate estimates of  $[\text{Fe}/\text{H}]$  for a typical star in the field. Thus, studies focused on metal-poor stars should adopt the synthetic-oversampling model, while studies examining the field should probably adopt the regression model.

Example *precision-recall* curves are shown in Figure 5. The *precision-recall* curves confirm what is shown in Table 2: models with synthetic oversampling perform better than the baseline model. In particular, the synthetically-oversampled models show comparable or dramatically improved *precision* for any *recall*  $\lesssim 0.25$ . Interestingly, the synthetic-oversampling method does not provide a significant boost relative to the baseline model for *recall*  $> 0.3$ . Figure 5 also shows that the differences between the optimized synthetic-oversampling models is small, as was suggested by Table 2. It is also worth noting that the majority of false positives for the synthetic-oversampling models are metal poor:  $\sim 65\%$  of the false positives are VMP stars. Finally, note that the SVM regression model presented in §4 does a good job of recovering VMP stars without any additional tuning (see the dashed line in Figure 5). The SVM regression model produces a *recall*  $\approx 0.55$  at a *precision* of  $\sim 0.5$ .

### 5.3. Potential Biases in the EMP Sample

If the EMP stars in the training set are not representative of the true distribution of EMP stars in the field, then the synthetic-oversampling method will produce a biased sample. It is further possible that synthetic oversampling preferentially selects a specific type of EMP star, such as cool dwarfs or hot sub-giants. If present, these biases would prevent the construction of a complete sample. The number of known EMP stars is small enough, however, that any additional discoveries are valuable for understanding these rare stars. Furthermore, the biases in these methods may be complementary to other methods. For instance, the infrared-color technique presented in Schlafman & Casey (2014) preferentially selects giant stars.

Examining which EMP stars are recovered via CV can provide an estimate of the bias in the synthetic-oversampling method. Figure 3 shows the CV-recovered EMP stars when using a candidate decision threshold of  $[\text{Fe}/\text{H}]_{\text{SOM}} \leq -2.707$ . The results shown are for the model with 4000% oversampling and no downsampling, though the total number, and location, of sources recovered does not change significantly for any of the models with FoM  $\approx 0.2$ . From Figure 3, it is clear that the EMP stars closest to the edges of the stellar locus are the most likely to be recovered. This is not surprising for two reasons: (1) there is higher contrast between

TABLE 2  
OPTIMIZED EMP CLASSIFICATION RESULTS

ds <sup>a</sup>	N <sup>b</sup>	ROC AUC	PR AUC	F-measure	R(P = 0.1) <sup>c</sup>	R(P = 0.05) <sup>d</sup>	P(R = 0.1) <sup>e</sup>
25	0	0.909	0.022	0.000	0.000	0.132	0.049
25	1000	0.916	0.027	0.064	0.022	0.197	0.052
25	2000	0.920	0.026	0.078	0.016	0.198	0.056
25	3000	0.921	0.025	0.074	0.007	0.193	0.052
25	4000	<b>0.922</b>	0.025	0.066	0.003	0.178	0.053
50	0	0.909	0.023	0.000	0.001	0.133	0.055
50	1000	0.915	0.027	0.044	0.009	0.198	0.055
50	2000	0.917	0.027	0.054	0.027	<b>0.212</b>	0.059
50	3000	0.919	0.026	0.074	0.003	<b>0.212</b>	0.058
50	4000	0.920	0.025	<b>0.084</b>	0.001	0.203	0.053
75	0	0.909	0.024	0.000	0.004	0.109	0.054
75	1000	0.913	0.024	0.000	0.000	0.203	0.054
75	2000	0.916	0.027	0.051	<b>0.033</b>	0.207	<b>0.060</b>
75	3000	0.917	0.028	0.066	0.029	0.201	0.056
75	4000	0.918	0.027	0.075	0.020	0.207	0.059
100	0	0.909	0.024	0.000	0.004	0.109	0.056
100	1000	0.912	0.024	0.000	0.000	0.207	0.055
100	2000	0.915	0.027	0.033	0.023	<b>0.212</b>	0.059
100	3000	0.916	0.027	0.050	0.025	0.210	0.055
100	4000	0.918	<b>0.028</b>	0.065	<b>0.033</b>	<b>0.212</b>	0.058

NOTE. — Bold quantities indicate the maximum for a given column. Table values represent the average of 3 different instances of 10-fold cross validation.

<sup>a</sup> Percentage of the majority class remaining following downsampling.

<sup>b</sup> Percentage increase in the minority class via synthetic oversampling (see text).

<sup>c</sup> The recall at precision = 0.1.

<sup>d</sup> The recall at precision = 0.05. This is the model FoM.

<sup>e</sup> The precision at recall = 0.1.

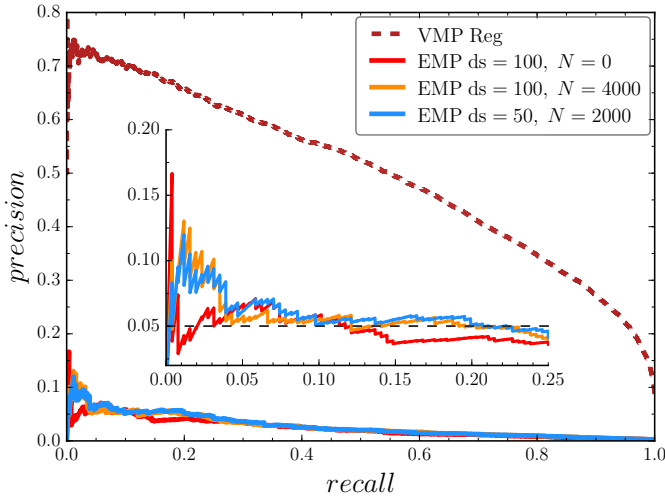


FIG. 5.— *precision-recall* curves for different parameters of the synthetic oversampling method. The baseline model, which features no oversampling ( $N = 0$ ) or downsampling of the majority ( $ds = 100$ ), is shown with a solid red line. The model with no downsampling and 4000% ( $N = 4000$ ) oversampling is shown with a solid orange line, while the model with 50% downsampling ( $ds = 50$ ) and 2000% ( $N = 2000$ ) oversampling is shown with the light blue line. The use of synthetic EMP stars significantly improves the performance of the model. The zoom-in shows that the synthetic-oversampling method produces a FoM that is  $\sim 2\times$  better than the baseline model. These results are relatively insensitive to the degree of over- and downsampling (see Table 2). The maroon-dashed line shows the *precision-recall* curve when using the SVM regression model from §4 to identify VMP stars.

these EMP stars and the background of non-EMP stars, and (2) the model has been optimized to sacrifice completeness in favor of *precision*. Examination of the SSPP parameters for the recovered EMP stars shows that the synthetic oversampling method is biased towards recov-

ering warm stars with relatively high surface gravity. In particular, of the 126 EMP stars in the training set with  $T_{\text{eff}} \geq 6000$  K,  $\sim 28\%$  are recovered, while only  $\sim 15\%$  of the 130 stars with  $T_{\text{eff}} < 6000$  K are recovered. Of the 173 stars with  $\log g < 3.5$  dex,  $\sim 14\%$  are recovered, while  $\sim 36\%$  of the 83 EMP stars with  $\log g \geq 3.5$  dex are recovered. Thus, it can be concluded that the synthetic oversampling model preferentially selects the hotter, higher surface gravity stars within our training set.

It is significantly more complicated to determine whether or not the training set is biased relative to the true population of EMP stars. This is primarily because the actual distribution of EMP stars is unknown, but the complex targeting procedures adopted by the SDSS-I and SDSS-II surveys further muddies the picture. Furthermore, the targeting criteria for the SEGUE portion of SDSS, which is responsible for most of the stellar spectra included in this study, evolved with time to improve the efficiency of target selection (Yanny et al. 2009). Spectroscopic targets were selected using a variety of cuts on brightness, photometric color, and proper motion to identify stars belonging to different classes, e.g., white dwarfs, K giants, G stars, etc. As a result, the population of EMP stars detected by SDSS must be biased. In particular, SEGUE used photometric metallicity indicators to preferentially select metal-poor (MP) and metal-poor, turn-off (MPTO) stars. 120 of the 256 EMP stars in the training set were targeted as either MP or MPTO stars. SEGUE biased their search for metal-poor stars toward hotter, and thus more luminous, stars, even though K and M dwarfs live much longer on the main sequence, in order to probe a larger effective survey volume (Yanny et al. 2009). As a result, there are few cool EMP stars in the SDSS spectroscopic sample (see the relative lack of EMP stars with  $g - r \gtrsim 0.6$  mag in Figure 3). While

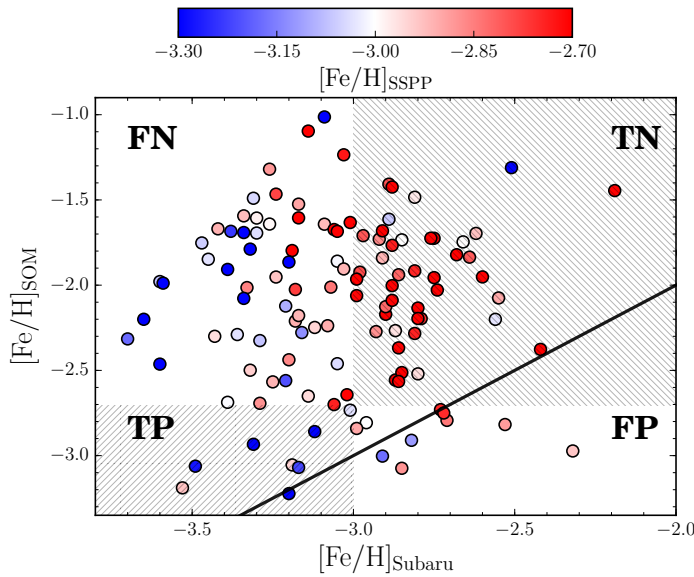


FIG. 6.— Results from the synthetic-oversampling LOO CV performed on the 119 stars in common between this study and the high-resolution spectroscopic sample of Aoki et al. (2013). The synthetic-oversampling method predicted  $[\text{Fe}/\text{H}]$ ,  $[\text{Fe}/\text{H}]_{\text{SOM}}$ , is shown relative to the Subaru/HDS measured  $[\text{Fe}/\text{H}]$ . Stars are color coded by their SSPP measured  $[\text{Fe}/\text{H}]$ . The solid, black line shows the relation for perfect 1:1 regression. All stars with predicted  $[\text{Fe}/\text{H}] \leq -2.707$  are considered EMP candidates. The areas corresponding to true positives and true negatives are shaded in the lower-left and upper-right corners, respectively. False negatives and false positives are shown in the upper-left and lower-right corners, respectively.

a quantitative measure of this bias is not available, it is clear that, by design, the sample of EMP stars identified by SDSS spectroscopy is biased towards F turnoff stars.

This training set bias provides context for understanding the biased selection of EMP stars from the synthetic-oversampling method. The over-representation of  $\sim$ F-type stars in the training set is naturally propagated through the machine-learning model to preferentially recover warm ( $T_{\text{eff}} \gtrsim 6000$  K) EMP stars.

#### 5.4. Confirmation of the Synthetic-Oversampling Method with High-Resolution Spectra

The best way to confirm the efficacy of the synthetic-oversampling method is to obtain spectra of candidate EMP stars, measure  $[\text{Fe}/\text{H}]$  for these stars, and determine whether or not the *precision* of the sample is  $\approx 0.05$ , as was predicted in §5.2. Current efforts to obtain such spectra are ongoing and the subject of a future study. In the meantime, the model accuracy can be tested using the high-resolution spectra obtained by Aoki et al. (2013). Using the High Dispersion Spectrograph (HDS) on the Subaru Telescope, Aoki et al. obtained high-resolution spectra of 137 candidate EMP stars selected from the SSPP. Relative to the SSPP, the HDS spectra provide more accurate and precise measurements of  $[\text{Fe}/\text{H}]$ , leading to the unambiguous identification of EMP stars. There are 119 stars in common between the Aoki et al. sample and this study. Leave-one-out (LOO) CV<sup>12</sup> is used to measure the fraction of the EMP stars

recovered by the synthetic-oversampling method.

Figure 6 shows the results of the LOO CV procedure, based on a model with 4000% oversampling and no down-sampling. The results show a significant improvement over those shown in Figure 2, where synthetic oversampling is not employed and there are no stars with predicted  $[\text{Fe}/\text{H}] \leq -2.5$  dex. Of the 119 stars, 64 are genuine EMP stars, and the synthetic-oversampling method identifies 8 of those as EMP candidates based on their photometric colors. This corresponds to *recall* = 0.125, which is worse than expectations (see Table 2). More promising is the paucity of false positives, 10, which corresponds to a *precision*  $\approx 0.44$ . This estimate of the *precision* is likely over-optimistic, however, because EMP stars outnumber VMP stars in the Aoki et al. (2013) sample. In the halo, VMP stars outnumber EMP stars by a factor of  $\sim 50$  (e.g., Allende Prieto et al. 2014). Pessimistically, this would suggest a ratio of  $\sim 500$  false positives for every  $\sim 8$  true positives, corresponding to a *precision*  $\approx 0.02$ . While this is not too dissimilar from the expected *precision* for the model, the true *precision* of the model is likely better than 0.02. The Aoki et al. sample of VMP stars is significantly skewed towards stars with  $[\text{Fe}/\text{H}] \leq -2.7$ , while the actual halo metallicity distribution strongly favors stars with  $[\text{Fe}/\text{H}] \approx -2.0$  relative to stars with  $[\text{Fe}/\text{H}] \leq -2.7$ . Assuming the model is less likely to identify the most metal-rich VMP stars as EMP candidates, which CV shows is the case, then the *precision* should be better than the pessimistic estimate.

Finally, note that the stars shown in Figure 6 are color coded by  $[\text{Fe}/\text{H}]_{\text{SSPP}}$ . While the *recall* is worse than one would expect based on the CV results from §5.2, it is worth noting that the SSPP slightly over-predicts  $[\text{Fe}/\text{H}]$ . In particular, only 36 of the 119 stars have  $[\text{Fe}/\text{H}]_{\text{SSPP}} \leq -3.0$ , meaning the SSPP sample is biased away from EMP stars, as determined by the Subaru spectra. Relative to the  $[\text{Fe}/\text{H}]_{\text{SSPP}}$  labels, the synthetic-oversampling model produces a *recall* of 0.25, as one would expect based on the results shown in Table 2. Ultimately, this is a demonstration that the results of the machine-learning models are only as good as the training set. When comparing the SSPP measurements to those from the high-resolution spectra, the SSPP has a *recall*  $\approx 0.47$ , assuming a class boundary of  $[\text{Fe}/\text{H}]_{\text{SSPP}} = -3.0$ . Assuming that the spectroscopic measurements from the Subaru spectra are more accurate than the SSPP, this means the synthetic-oversampling method has a *recall* ceiling of  $\sim 0.47$ . Moving forward, there are two paths towards improving this ceiling: (i) improve the accuracy of the SSPP measurements, or (ii) obtain significantly more high-resolution spectra, and build a model using those  $[\text{Fe}/\text{H}]$  measurements. While several incremental improvements have been made to the SSPP (e.g., Ahn et al. 2012; Schlesinger et al. 2012; Aoki et al. 2013), low-resolution spectra will always produce lower-accuracy measurements than their high-resolution counterparts. Furthermore, high-resolution spectra are extremely expensive, meaning a new, uniformly analyzed training set with  $> 10^5$  sources is unlikely to be available any time soon.<sup>13</sup> Thus, for the foreseeable future, and despite

the left-out star. For this study, this procedure is repeated for each of the 119 stars with HDS observations.

<sup>13</sup> SDSS has obtained a set of high-resolution near-infrared spec-

<sup>12</sup> LOO CV is similar to  $k$ -fold CV. The difference is that rather than testing on all the data, LOO CV removes a single star from the training set, constructs a model, and then predicts  $[\text{Fe}/\text{H}]$  for



some clear limitations, the SSPP provides the best basis for a training set to search for EMP stars.

## 6. FINAL FIELD-STAR PREDICTIONS

Finally,  $[\text{Fe}/\text{H}]$  values are predicted for all SDSS stars that satisfy the same selection criteria as the training set (see §2). In sum, there are 14,337,770 sources in SDSS DR10 that satisfy all of those photometric criteria, and have  $\text{ProfPSF} = 1$ , which excludes sources with extended morphologies. Predicted  $[\text{Fe}/\text{H}]$  values, from both the SVM-regression model (see §4.1) and the synthetic-oversampling model (see §5), for most of these stars are reported in Table 3, though important caveats apply.

The first caveat is that, unlike for the training set, this photometric sample does not have spectroscopic measurements of  $T_{\text{eff}}$ . Given that the training set only includes stars satisfying  $4500\text{K} \leq T_{\text{eff}} \leq 7000\text{K}$ , the machine-learning models will not produce reliable predictions for stars outside this temperature range. To select stars that satisfy this criteria,  $T_{\text{eff}}$  is assigned to the photometric sample using the Color- $T_{\text{eff}}$  relations in Pinsonneault et al. (2012). These Color- $T_{\text{eff}}$  relations are calibrated for  $4080\text{ K} \leq T_{\text{eff}} < 7000\text{ K}$ , which covers the full range of  $T_{\text{eff}}$  included in the training set. As their method is not valid at all temperatures, Pinsonneault et al. caution that the three individual relations are only valid for stars with  $0.13 < (g-r)_0 < 1.34$ ,  $0.13 < (g-i)_0 < 1.90$ , and  $0.07 < (g-z)_0 < 2.21$ , respectively. Stars with colors outside this range are excluded from Table 3, which restricts the sample of field stars to 13,004,005. The three Color- $T_{\text{eff}}$  relations, one each for  $(g-r)_0$ ,  $(g-i)_0$ , and  $(g-z)_0$ , are applied to each star and the mean  $T_{\text{eff}}$  is adopted. There are 12,735,277 stars with a mean  $T_{\text{eff}}$  between 4500 K and 7000 K, and they are summarized in Table 3.<sup>14</sup>

The second caveat is that most data-driven methods are not reliable outside the parameter space enclosed by the training set. Figure 1 shows the training set is confined to a specific location in feature space, i.e. the stellar locus. Thus, model predictions for sources within the range of acceptable  $T_{\text{eff}}$ , but well outside the region defined by the training set, may be unreliable. To aid the user in identifying potentially unreliable estimates of  $[\text{Fe}/\text{H}]$ , Table 3 includes a proximity measure  $\rho$ , which measures the relative distance of any given star to the

training set. The proximity measure is defined as:

$$\rho_i = \frac{1}{60} \sum_{j=1}^{60} \left[ \sum_{l=1}^4 (x_{i,l} - x_{j,l})^2 \right]^{1/2}, \quad (2)$$

where  $\rho_i$  is the proximity measure of the  $i^{\text{th}}$  source,  $x_i$  is the 4-dimensional feature vector, with each feature,  $l$ , scaled to the standard normal distribution, respectively, and the sum is over each of the  $j$  60-nearest neighbors to the  $i^{\text{th}}$  source as determined by the KNN algorithm. Thus,  $\rho$  represents the mean Euclidean distance between a given source and its 60-nearest-training-set neighbors.<sup>15</sup> Sources with large  $\rho$  are likely to have unreliable estimates of  $[\text{Fe}/\text{H}]$ .

The proximity measures are relative, meaning there is no hard and fast rule for a threshold on  $\rho$  that eliminates all unreliable  $[\text{Fe}/\text{H}]$  estimates. Table 4 shows the proximity measure for training set sources based on several commonly adopted threshold percentiles. Studies that require high-fidelity  $[\text{Fe}/\text{H}]$  estimates can adopt a small  $\rho$  threshold, while studies requiring larger samples can relax that criterion. Figure 7 shows training-set and field stars that would be considered unreliable when adopting a proximity-measure threshold of  $\rho_t = 0.3883$ , corresponding to the most distant 1% training-set stars. The top panel of Figure 7, which highlights sources in the training set, shows that nearly every training-set source outside the 99.7%  $(u-g)_0$ ,  $(g-r)_0$  contour is flagged as unreliable. Sources with  $\rho > 0.3883$  inside the contour have anomalous  $(g-i)_0$  or  $(g-z)_0$  colors. Applying the same threshold to the field stars in Table 3, shown in the bottom panel of Figure 7, shows that sources distant from the training set are flagged as unreliable. In particular, once again the vast majority of stars outside the 99.7% contour are flagged as unreliable. It is reassuring that the cluster of sources located at  $(u-g)_0$ ,  $(g-r)_0 \approx (0.15, 0.2)$ , which should be dominated by quasars (Sesar et al. 2007), is flagged with large  $\rho$ . Of the 12,735,277 field stars to which the model is being applied, the  $\rho_t = 0.3883$  threshold would flag  $\sim 1.7\%$  as potentially unreliable. That this number is close to 1% suggests that the distribution of stars in the training set and the field are very similar.

The synthetic-oversampling predictions ( $[\text{Fe}/\text{H}]_{\text{SOM}}$ ) presented in Table 3 come from the model with 4000% oversampling and no downsampling. When using this model, any stars with  $[\text{Fe}/\text{H}]_{\text{SOM}} \leq -2.707$  are considered EMP candidates. Adopting this threshold results in 17,605 candidates in Table 3. That threshold corresponds to a *precision* = 0.05, meaning  $\sim 880$  of these candidates should be genuine EMP stars. This estimate ignores the proximity measure, however, and thus likely overestimates the true number of EMP stars in the sample. The application of a conservation proximity-measure threshold,  $\rho \leq 0.1705$ , which corresponds to the 95<sup>th</sup> percentile of the training set (see Table 4), reduces the sample to 11,491,213 stars with 11,849 EMP candidates. Of these candidates,  $\sim 590$  should be bonafide EMP stars given the *precision* of the synthetic-oversampling model.

tra that is this large (Alam et al. 2015), however, that sample includes very few VMP stars and virtually no EMP stars.

<sup>14</sup> The color- $T_{\text{eff}}$  relations presented in Pinsonneault et al. (2012) are calibrated for dwarf stars at  $[\text{Fe}/\text{H}] = -0.2$  dex. A change in  $[\text{Fe}/\text{H}]$  results in a change in  $T_{\text{eff}}$  at fixed color. As a result the limits placed on the photometrically-determined  $T_{\text{eff}}$  will slightly bias the sample towards metal-poor stars by including some that are cooler than 4500 K and hotter than 7000 K (see Table 3 in Pinsonneault et al. 2012). Given the magnitude limits on the sample and the rarity of metal-poor stars, the overall contamination is expected to be very small. The color- $T_{\text{eff}}$  relations have a weak dependence on  $\log g$ , with only cool, giants ( $T_{\text{eff}} \lesssim 5000$  K;  $\log g \lesssim 3.5$ ) requiring corrections. Stars with  $\log g \approx 2.0$  need the most significant corrections, which, nevertheless, are relatively small ( $\Delta T_{\text{eff}} \lesssim 100$  K). Given the rarity of giants in the SDSS photometric sample (see e.g., Ivezić et al. 2008b), these corrections are ignored and should not significantly bias the final predictions from the models.

<sup>15</sup> The choice of 60 neighbors is arbitrary, but the relative ranking of the proximity measure does not change significantly for any choice of  $k \gg 1$  neighbors.  $k = 60$  was adopted to match the optimized KNN model from §4.1.

TABLE 3  
FINAL METALLICITY PREDICTIONS FOR FIELD STARS

Name	Object ID <sup>a</sup>	$\alpha_{J2000.0}$ (hh:mm:ss.ss)	$\delta_{J2000.0}$ (dd:mm:ss.s)	$T_{\text{eff}}^b$ (K)	[Fe/H] <sub>SVM</sub> <sup>c</sup> (dex)	[Fe/H] <sub>SOM</sub> <sup>d</sup> (dex)	$\rho^e$
SDSS J000000.00+204152.5	1237680247351279746	00:00:00.00	+20:41:52.5	5617	−2.024	−2.425	0.0963
SDSS J000000.01+345915.4	1237666184574271704	00:00:00.01	+34:59:15.4	4579	−0.568	−0.567	0.1641
SDSS J000000.02+125954.1	1237678920204681228	00:00:00.02	+12:59:54.1	5903	−0.815	−0.828	0.0627
SDSS J000000.03+032107.2	1237678620102164731	00:00:00.03	+03:21:07.2	5340	−0.581	−0.587	0.0898
SDSS J000000.04+015313.0	1237678596479844501	00:00:00.04	+01:53:13.0	4726	−0.307	−0.311	0.1030
SDSS J000000.05−005019.4	1237663783123681350	00:00:00.05	−00:50:19.4	6110	−0.960	−0.958	0.0466
SDSS J000000.05+065743.2	1237669680114106516	00:00:00.05	+06:57:43.2	5010	−0.781	−0.779	0.0830
SDSS J000000.07+333115.1	1237663307989909606	00:00:00.07	+33:31:15.1	5631	−0.394	−0.393	0.0608
SDSS J000000.08+202502.3	1237679504318922768	00:00:00.08	+20:25:02.3	5467	0.166	0.162	0.1057
SDSS J000000.08+305810.6	1237663234451309002	00:00:00.08	+30:58:10.6	5148	−0.356	−0.348	0.0842

NOTE. — Only the first ten sources are presented here as an example of the form and content of the complete table. The full table, containing all 12,735,277 SDSS point sources with [Fe/H] predictions, is available online.

<sup>a</sup> objID from the SDSS DR10 PhotoObjAll table.

<sup>b</sup> Photometrically determined  $T_{\text{eff}}$  using the method of Pinsonneault et al. (2012). See text for further details.

<sup>c</sup> Photometric [Fe/H] determined using the SVM-regression model from §4.1.

<sup>d</sup> Photometric [Fe/H] determined using the synthetic-oversampling method (see §5). Note – stars with [Fe/H]<sub>SOM</sub> ≤ −2.707 are EMP candidates.

<sup>e</sup> The proximity measure,  $\rho$ . See Table 4 for useful thresholds on  $\rho$ .

TABLE 4  
PROXIMITY MEASURE THRESHOLDS

Percentile	$\rho_t$
68	0.0843
90	0.1310
95	0.1705
99	0.3883
99.5	0.5774
99.7	0.7737

NOTE. — The threshold,  $\rho_t$ , corresponding to the percentage of training set sources with  $\rho \leq \rho_t$ .

There are only a few hundred known EMP stars that have been confirmed with high-resolution spectra (e.g., Aoki et al. 2013; Roederer et al. 2014; Jacobson et al. 2015), the discovery of ~600 new members of the class would represent a huge windfall for this field of study.

## 7. SUMMARY AND CONCLUSIONS

I have presented a new photometric method for inferring stellar metallicity from the SDSS *ugriz* filters. The model, which utilizes machine-learning algorithms, is capable of identifying previously unknown EMP stars once the training set has been supplemented with synthetic EMP stars. The model is trained using a large sample of SDSS stars with high SNR spectra and reliable measurements of [Fe/H] from the SSPP. Following reasonable cuts on photometric and spectroscopic quality, and the removal of duplicate spectra of the same star, the training set consists of 170,610 unique stars.

The [Fe/H] regression model represents an improvement over previous methods by utilizing all four non-redundant colors and a non-parametric model capable of capturing complex interactions between the colors. Three separate models, *k*-nearest neighbors, random forest, and support vector machines, are trained, and optimized, using 110,000 stars from the training set. When these models are applied to the 60,610 star independent validation set, they each produce an RMSE  $\approx 0.29$  dex relative to the spectroscopic measurements of [Fe/H]. Of

the three models, SVM produces the smallest RMSE and bias, though the improvement relative to *k*NN and RF is small ( $\lesssim 1\%$ ). The performance of the machine-learning models is compared to that of low-resolution spectra, which produce a typical scatter of  $\sim 0.24$  dex when measuring [Fe/H] (Lee et al. 2008a). Using a sample of stars with high-resolution spectroscopic observations as ground truth, the SSPP provides only a  $\sim 17\%$  improvement over the photometric method presented in this paper. Thus, the [Fe/H] regression methods presented here are comparable to the accuracy achieved with low-resolution spectra, with the major benefit that photometric colors can be acquired much cheaper than spectra. Furthermore, it was demonstrated that the machine-learning regression methods perform better than other photometric [Fe/H] techniques, while also being more general. In particular, there is an  $\sim 18\%$  improvement relative to the methods presented in Ivezić et al. (2008b) and Bond et al. (2010). As a demonstration of the fidelity of the model, [Fe/H] predictions for 12,735,277 stars without spectroscopic observations are presented in Table 3. Proximity measures are provided for the  $\sim 12$  million stars with [Fe/H] predictions, in order to evaluate the reliability of the individual estimates.

A challenge for this method, and all photometric-metallicity techniques, is correcting for interstellar reddening. The ability to measure [Fe/H] directly from absorption lines, independent of reddening, remains a major advantage of spectroscopy. In principle, data-driven photometric methods could be used to recover  $T_{\text{eff}}$ , [Fe/H], and extinction (the method presented in §4 effectively recovers  $T_{\text{eff}}$  and [Fe/H]), but that would require a significantly enhanced training set. Typically, a training set must grow by  $\sim$ an order of magnitude to properly capture the diversity necessary to resolve a new parameter. Even with such an expanded training set, I speculate that broadband filters will struggle to fully break the degeneracies between these parameters (unless there is also a significant improvement in photometric precision). Thus, broadband photometric-metallicity techniques are, and will remain, limited in the vicinity of

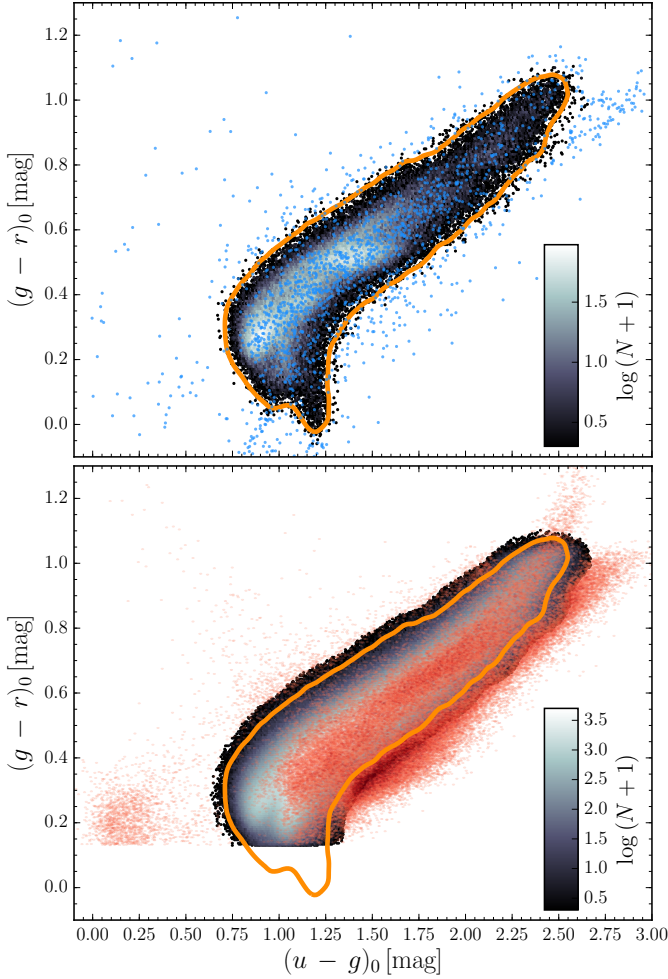


FIG. 7.—  $(u - g)_0$ ,  $(g - r)_0$  CC diagram showing the location of stars with large proximity measure. *Top*: density plot showing the total number of *training set* stars in each  $\sim 0.01 \times 0.01$  mag pixel on a white to black color scale. The blue points show training set stars in the 99<sup>th</sup> percentile of proximity measure, corresponding to  $\rho \geq 0.3883$ . The solid orange line shows the 99.7% contour for the training set, as measured in the  $(u - g)_0$ ,  $(g - r)_0$  plane. *Bottom*: density plot showing the total number of *field stars* with  $\rho < 0.3883$  per pixel. The solid orange line shows the same contour as the top panel. Red points show the location of field stars with  $\rho \geq 0.3883$ . The majority of stars outside the main stellar locus have large proximity measure. There are no field stars with  $(g - r)_0 \lesssim 0.13$ , because the Colors- $T_{\text{eff}}$  method does not apply to stars with  $T_{\text{eff}} > 7000$  K.

the Galactic plane.

A primary aim of developing the photometric model was to discover EMP stars. There is a significant class imbalance in the training set,  $< 0.2\%$  of the sample consists of EMP stars, making it difficult to identify these rare relics of the early universe. To improve the recoverability of these sources, a new framework, referred to as the synthetic-oversampling method, was developed where synthetic EMP are added to the training set while a randomly selected fraction of the majority (non-EMP) class stars are removed.

The goal of the synthetic-oversampling method is to identify EMP stars, while having a relatively low tolerance for false positives. Thus, the adopted FoM is to maximize the model *recall* at a fixed *precision* = 0.05, which corresponds to 19 false positives for every newly

discovered EMP star. It is found that the synthetic-oversampling method outperforms the baseline model, where no oversampling or downsampling have occurred, with a *recall*  $\approx 0.2$  at *precision* = 0.05. This represents a  $\sim 100\%$  increase in the FoM relative to the baseline model. The synthetic oversampling method was further tested using 119 stars with high-resolution spectroscopic observations from Aoki et al. (2013). This sample includes 64 bonafide EMP stars, and the use of leave-one-out CV shows that the model produces a *recall*  $\approx 0.125$ .

An examination of the EMP stars that are recovered by the synthetic-oversampling method shows that there is a bias towards the selection of warm ( $T_{\text{eff}} \gtrsim 6000$  K) stars with relatively high surface gravities ( $\log g \gtrsim 3.5$ ). The SEGUE target selection of metal-poor stars was intentionally biased towards warmer stars (Yanny et al. 2009), which probe larger volumes at fixed mag, which, in turn leads to a bias in the training set for this study. Thus, the bias introduced by SEGUE is propagated through to the synthetic-oversampling method.

While 19 false positives for every EMP star seems high, this represents a significant improvement over the metal-poor candidate selection techniques adopted by SEGUE. In particular, within the training set 20,200 stars were targeted as metal-poor, and only  $\sim 0.3\%$  are EMP stars. Another 18,606 were targeted as likely metal-poor turnoff stars, with, again, a yield of only  $\sim 0.3\%$ . Furthermore, of the 19 false positives for every EMP star,  $\sim 65\%$  of those false positives are VMP stars. Thus, the synthetic-oversampling method produces a highly pure sample of metal-poor stars. Future and ongoing spectroscopic surveys hoping to efficiently identify large samples of EMP stars, such as the Large Sky Area Multi-Object fiber Spectroscopic Telescope (LAMOST; Cui et al. 2012; Deng et al. 2012), should adopt the synthetic-oversampling method for target selection.

Astronomy has embarked upon an age where wide-field photometry is cheap: SDSS and Pan-STARRS (Kaiser et al. 2010) have mapped a significant fraction of the sky in multiple filters to a depth of  $\sim 21$ -22 mag. LSST will do the same for the entire southern sky to a depth of  $\sim 27$  mag. Now, more than ever, it is imperative that meaningful physical information, such as  $[\text{Fe}/\text{H}]$ , can be extracted from photometric-only surveys. The large volume of data produced by LSST will prove no better than existing observations if the proper algorithmic solutions are not developed to deal with the new, complex data stream. Machine-learning methods provide a promising way to cope with the coming data deluge, and this work serves as a step in that direction. A great deal can be learned about the Galaxy from photometric metallicity measurements (e.g., Ivezić et al. 2008b; Bond et al. 2010), while the heaps of yet to be discovered EMP stars provide the promise of shedding light on otherwise unobservable aspects of the early universe.

I thank B. Bue and U. Rebbapragada for multiple useful conversations on class imbalance and model optimization. I am grateful that J. Cohen was willing to suffer many (possibly naive) questions about stellar metallicity measurements and bias in the SDSS sample. J. Cohen, L. Hillenbrand, and E. Kirby provided comments on an early version of this paper, which greatly improved its final content. Finally, I thank the SEGUE team for mak-



ing the results of the SSPP public, and I am especially indebted to Y. S. Lee, who has answered many inquiries about the SSPP flags and reliability of the individual [Fe/H] measurement methods.

I acknowledge support for this work by NASA from a Hubble Fellowship grant: HST-HF-51325.01, awarded by STScI, operated by AURA, Inc., for NASA, under contract NAS 5-26555. Part of the research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is <http://www.sdss3.org/>.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Ari-

zona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofísica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

*Facilities:* Sloan

©2015. All rights reserved.

## REFERENCES

- Ahn, C. P., Alexandroff, R., Allende Prieto, C., et al. 2012, *ApJS*, **203**, 21
- . 2014, *ApJS*, **211**, 17
- Alam, S., Albareti, F. D., Allende Prieto, C., et al. 2015, ArXiv e-prints, [arXiv:1501.00963](https://arxiv.org/abs/1501.00963) [astro-ph.IM]
- Allende Prieto, C., Sivarani, T., Beers, T. C., et al. 2008, *AJ*, **136**, 2070
- Allende Prieto, C., Fernández-Alvar, E., Schlesinger, K. J., et al. 2014, *A&A*, **568**, A7
- An, D., Johnson, J. A., Beers, T. C., et al. 2009, *ApJ*, **707**, L64
- Aoki, W., Beers, T. C., Lee, Y. S., et al. 2013, *AJ*, **145**, 13
- Beers, T. C., & Christlieb, N. 2005, *ARA&A*, **43**, 531
- Beers, T. C., Preston, G. W., & Shectman, S. A. 1985, *AJ*, **90**, 2089
- . 1992, *AJ*, **103**, 1987
- Bond, N. A., Ivezić, Ž., Sesar, B., et al. 2010, *ApJ*, **716**, 1
- Boser, B. E., Guyon, I. M., & Vapnik, V. N. 1992, in Proceedings of the fifth annual workshop on Computational learning theory, ACM, 144
- Breiman, L. 1996, *Machine Learning*, **24**, 123
- . 2001, *Machine Learning*, **45**, 5
- Brink, H., Richards, J. W., Poznanski, D., et al. 2013, *MNRAS*, **435**, 1047
- Chang, C.-C., & Lin, C.-J. 2011, ACM Transactions on Intelligent Systems and Technology, **2**, 27:1, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002, *Journal of Artificial Intelligence Research*, **16**, 321
- Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. 2003, in Knowledge Discovery in Databases: PKDD 2003 (Springer), 107
- Chen, C., Liaw, A., & Breiman, L. 2004, Technical Report 666, Statistics Department, University of California, Berkeley
- Christlieb, N., Schörck, T., Frebel, A., et al. 2008, *A&A*, **484**, 721
- Cohen, J. G., Christlieb, N., McWilliam, A., et al. 2004, *ApJ*, **612**, 1107
- Cortes, C., & Vapnik, V. 1995, *Machine learning*, **20**, 273
- Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, *Research in Astronomy and Astrophysics*, **12**, 1197
- Deng, L.-C., Newberg, H. J., Liu, C., et al. 2012, *Research in Astronomy and Astrophysics*, **12**, 735
- Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. 1997, Advances in neural information processing systems, **9**, 155
- Dubath, P., Rimoldini, L., Süveges, M., et al. 2011, *MNRAS*, **414**, 2602
- Frebel, A., Christlieb, N., Norris, J. E., et al. 2006, *ApJ*, **652**, 1585
- Frebel, A. L., & Norris, J. 2015, *Annual Review of Astronomy and Astrophysics*, **53**, null
- Frost, C., & Thompson, S. G. 2000, *Journal of the Royal Statistical Society*, **163**, 173
- Hastie, T., Tibshirani, R., & Friedman, J. 2009, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer Series in Statistics (Springer)
- Ivezić, Ž., Tyson, J. A., Acosta, E., et al. 2008a, ArXiv e-prints, [arXiv:0805.2366](https://arxiv.org/abs/0805.2366)
- Ivezić, Ž., Sesar, B., Jurić, M., et al. 2008b, *ApJ*, **684**, 287
- Jacobson, H. R., Keller, S., Frebel, A., et al. 2015, ArXiv e-prints, [arXiv:1504.03344](https://arxiv.org/abs/1504.03344) [astro-ph.SR]
- Kaiser, N., Burgett, W., Chambers, K., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7733
- Keller, S. C., Schmidt, B. P., Bessell, M. S., et al. 2007, *PASA*, **24**, 1
- Keller, S. C., Bessell, M. S., Frebel, A., et al. 2014, *Nature*, **506**, 463
- Kerekes, G., Csabai, I., Dobos, L., & Trencsényi, M. 2013, *Astronomische Nachrichten*, **334**, 1012
- Lee, Y. S., Beers, T. C., Sivarani, T., et al. 2008a, *AJ*, **136**, 2022
- . 2008b, *AJ*, **136**, 2050
- Lianou, S., Grebel, E. K., & Koch, A. 2011, *A&A*, **531**, A152
- Ling, C. X., & Li, C. 1998, KDD, **98**, 73
- Miller, A. A., Bloom, J. S., Richards, J. W., et al. 2015, *ApJ*, **798**, 122
- Nordström, B., Andersen, J., Holmberg, J., et al. 2004, *PASA*, **21**, 129
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *Journal of Machine Learning Research*, **12**, 2825
- Pinsonneault, M. H., An, D., Molenda-Žakowicz, J., et al. 2012, *ApJS*, **199**, 30
- Richards, J. W., Starr, D. L., Butler, N. R., et al. 2011, *ApJ*, **733**, 10
- Richards, J. W., Starr, D. L., Brink, H., et al. 2012, *ApJ*, **744**, 192
- Roederer, I. U., Preston, G. W., Thompson, I. B., et al. 2014, *AJ*, **147**, 136
- Schlafly, E. F., & Finkbeiner, D. P. 2011, *ApJ*, **737**, 103
- Schlaufman, K. C., & Casey, A. R. 2014, *ApJ*, **797**, 13
- Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, *ApJ*, **500**, 525
- Schlesinger, K. J., Johnson, J. A., Rockosi, C. M., et al. 2012, *ApJ*, **761**, 160
- Schwarzschild, M., Searle, L., & Howard, R. 1955, *ApJ*, **122**, 353
- Sesar, B., Ivezić, Ž., Lupton, R. H., et al. 2007, *AJ*, **134**, 2236
- Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, *AJ*, **131**, 1163
- Soderblom, D. R. 2010, *ARA&A*, **48**, 581
- Strömgren, B. 1966, *ARA&A*, **4**, 433
- Wallerstein, G. 1962, *ApJS*, **6**, 407
- Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, *AJ*, **140**, 1868
- Yanny, B., Rockosi, C., Newberg, H. J., et al. 2009, *AJ*, **137**, 4377
- York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, *AJ*, **120**, 1579